Metacognition Learning (2017) 12:21–43
DOI 10.1007/s11409-016-9159-5

The impact of an online tool for monitoring and regulating learning at university: overconfidence, learning strategy, and personality

Anique B. H. de Bruin[1] · **Ellen M. Kok**[1] · **Jill Lobbestael**[2] · **Andries de Grip**[3]

Received: 20 October 2015 / Accepted: 16 May 2016 / Published online: 2 June 2016

Abstract Being overconfident when estimating scores for an upcoming exam is a widespread phenomenon in higher education and presents threats to self-regulated learning and academic performance. The present study sought to investigate how overconfidence and poor monitoring accuracy vary over the length of a college course, and how an intervention consisting of (1) a monitoring exercise and (2) a monitoring and regulation strategy, improves students' monitoring accuracy and academic performance. Moreover, we investigated how personality factors (i.e., grandiose and vulnerable narcissism, optimism) influence monitoring accuracy. We found that the Monitoring and Regulation Strategy positively influenced monitoring accuracy and exam scores, whereas the Monitoring Exercise that confronted students with their overconfidence protected students against overconfidence in the second exam score prediction but did not affect exam score. The results further revealed that exam score predictions lowered from the start to the end of the course for both poor and high performing students, but still leaving poor performers overconfident and high performers underconfident. Topic knowledge gained in the course did not wash out the Dunning Kruger effect, and results indicate that poor and high performers use different cues when predicting exam scores. Both grandiose and vulnerable narcissism contributed to overconfidence on exam score predictions but not on the Monitoring Exercise. These findings underline the potential of the Monitoring and Regulation Strategy intervention and ask for upscaling it to include measurements of self-regulated learning activities.

✉ Anique B. H. de Bruin
anique.debruin@maastrichtuniversity.nl

[1] School of Health Professions Education, Department of Educational Development and Research, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[2] Faculty of Psychology and Neuroscience, Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands

[3] Research Centre for Education and the Labour Market, School of Business and Economics, Maastricht University, Maastricht, The Netherlands

Keywords Monitoring · Regulation of learning · Absolute accuracy · Overconfidence · Online tool

When students are asked to predict their own performance, these predictions are often not in line with objective measures of that performance (Dunning et al. 2004). Especially poor performing students predict performing higher than warranted based on objective outcomes, sometimes up to 30% (Hacker et al. 2000). This overconfidence stands in the way of accurate self-regulated learning (SRL) and academic achievement, and as such contributes to failure rates in higher education (e.g., Bol et al. 2005; Miller and Geraci 2011a). This intricate relationship between accurate monitoring of learning, self-regulated learning, and learning outcomes is central to many models of SRL (e.g., Pintrich et al. 2000; Nelson and Narens 1990; Zimmerman 2000). Scrutinizing potential causes of poor monitoring and how interventions can be designed to improve this has been subject of much research over the last two decades (Dunlosky et al. 2011; Rawson and Dunlosky 2007; Thiede et al. 2003).

In this study, we examined (1) how students' monitoring as measured through predictions of exam scores changed during the length of a course, and (2) how feedback on their monitoring accuracy and a strategy to improve SRL (through an online intervention) influenced their monitoring accuracy and exam grades. We further examined (3) to what extent overconfidence is related to certain personality traits. We will first summarize recent research on overconfidence in exam score predictions, then detail how interventions might influence monitoring accuracy, and finally describe how personality factors potentially contribute to overconfidence.

Explaining overconfidence in higher education

Overconfidence in predicting exam grades is particularly common among poor performing students (Bol et al. 2005; Hacker et al. 2000; Kelemen et al. 2007; Kruger and Dunning 1999; Miller and Geraci 2011a; Nietfeld et al. 2006), a phenomenon termed the Dunning-Kruger effect. In contrast, high performing students typically provide accurate predictions or are even underconfident. Although evidence regarding the direction of causality between these variables is difficult to obtain, explanations that have been postulated for poor performers' overconfidence include describing it as a measurement artifact: Regardless of actual exam scores, students' predictions all hoover around 70-80%, which is close to high performers scores, but off the mark for poor performers (Krueger and Mueller 2002). In addition, Hartwig and Dunlosky (2014) provide evidence that overconfidence is mainly observed when students produce percentile rank judgments (i.e., when students had to judge their performance relative to other students performance as a percentage) and less or even absent for absolute grade judgments. They interpreted students' difficulty generating accurate percentile rank judgments but not absolute grade judgments as indicating that students have difficulty judging *other* students' learning levels, which suggests that overconfidence is not necessarily an indication of poor *self*-awareness. Test item difficulty is also known to influence over- or underconfidence at the question item level. For difficult items, overconfidence is typically observed, whereas for easy items underconfidence is common. This robust observation, typically found within-judges, is termed the hard-easy effect (Lichtenstein et al. 1982) and explained by several factors, among which a biased selection of test questions (Gigerenzer et al. 1991), and students' insensitivity to actual difficulty of the items (Price 1998).

The most dominant, yet empirically underexplored hypothesis is the 'double curse' explanation (Kruger and Dunning 1999): Poor performing students not only lack sufficient knowledge, but also suffer from a deficit in metacognitive ability, which renders them unaware of their low knowledge level. Miller and Geraci (2011b) set out to examine this postulation by analyzing the *confidence* that students hold in their exam score predictions. If a low metacognitive ability explains poor performers' overconfidence, these students are expected to hold high confidence in their predictions. Again, poor performing students were overconfident in their grade predictions, but they were *less* confident about the accuracy of their predictions than were high performing students. These findings indicate at least partial awareness of their lack of metacognitive knowledge and argue against full metacognitive 'blindness' as suggested by the double curse account (see Händel and Fritzsche 2015, for research that shows that accuracy of confidence in confidence judgments (or second-order judgments) is higher in high-performing students). Miller and Geraci (2011b) also compared pre- to postdictions of performance, under the assumption that postdictions would be less prone to overestimation (Pierce and Smith 2001). Indeed, performance predictions lowered from pre- to post-test, to a similar extent for poor and high performing students. One could conclude that having experienced taking the test could protect poor performing students against overconfidence.

Now that we know that poor performing students are not fully metacognitively blind, it is time to further unearth the other side of the double curse; the role (a lack of) knowledge plays. The accuracy of students' metacognitive judgments is known to depend on the availability of diagnostic cues (Koriat 1997). These cues are either information-based or experience-based (Koriat et al. 2008). Information-based cues relate to beliefs or knowledge the student has about her competence and cognitions, e.g., the extent to which the student feels competent in the test subject. Experience-based cues originate from "sheer subjective feelings" (Koriat et al. 2008) and therefore only influence *post*-dictions of performance. These relate, e.g., to the speed of retrieving information from memory. When students gain more knowledge, this will increase their feeling of competence, thereby improving the diagnosticity of the cues they use when judging their learning. This in turn could increase their monitoring accuracy. Knowledge gain thereby influences information-based cues. Knowledge gain could also influence experience-based cues, because fluency of information retrieval improves. By having students predict their final exam score at the beginning of a college course and once again close to the exam, the influence of knowledge on exam score predictions can be studied. If availability of knowledge influences experience- and information-based cues, and thus monitoring accuracy, overconfidence will decrease over the length of the course as knowledge increases. By improving cue use, knowledge gain potentially also protects against underconfidence. If knowledge does not play a role, overconfidence will remain stable. Differentiating between poor and high performers in these analyses will shed light on the knowledge versus metacognitive ability debate. Knowledge will increase more during the course for high versus poor performers, whereas metacognitive ability will remain stable. If availability of knowledge is paramount to accurate monitoring, high performers' grade predictions will become more accurate (i.e., less over- and underconfidence) than poor performers' predictions.

Interventions to improve exam score predictions

If we desire to improve students' predictions of their exam scores, a crucial intermediate step is to aid them in improving their ongoing monitoring during individual study sessions. Moreover,

Nietfeld et al. (2006) provided evidence that repeated monitoring exercises during the length of a college course can affect exam performance. Comparable to Hacker et al. (2000), they called for more classroom-based studies attempting to measure and improve students' monitoring accuracy. During the length of a 16-week college course, students completed monitoring exercises in class by self-assessing both their knowledge and preparation at the end of each class. Moreover, they provided confidence judgments for four intermediate tests and for review items during class. They were specifically instructed to reflect on the accuracy of their test responses in relation to their confidence judgments. Compared to a group of students who only provided confidence judgments on the intermediate tests, students who performed the monitoring exercises improved the accuracy of their judgments (i.e., the absolute difference between confidence judgments and test performance) over the length of the course, but also their test scores. Both improvements were clearly related: those who improved their judgments more also increased their test score more. In a study by Kleitman and Costa (2014) students had the opportunity to take 9 quizzes (formative assessment) during a statistics course. These quizzes contained multiple-choice questions on the course topic, but also asked for confidence judgments and provided feedback on confidence accuracy. Results showed that those who attempted more quizzes, and those who were more accurate in exam predictions performed higher on the final course exam. This effect was strongest for the poor performers. In a similar vein, Miller and Geraci (2011a) examined the effect of feedback on four exam score predictions and noticed that particularly poor performing students benefited by reducing their overconfidence. However, their exam scores did not improve. Miller and Geraci postulated that further instructions on how to improve monitoring judgments, for instance through self-testing, might broaden the scope of classroom interventions from merely monitoring to monitoring as well as control. These findings emphasize the positive effect monitoring exercises during class can have and warrant the use of classroom interventions.

In line with the abovementioned studies, we argue that accurate predictions of one's exam scores crucially hinge on continuous monitoring of learning during everyday self-regulated study. But how can we optimize monitoring of everyday study activities outside the classroom? And how can we design interventions that not only improve monitoring but also affect control to solicit the desired effect on academic achievement? Contrary to Nietfeld et al.'s (2006) and Miller and Geraci's (2011b) classroom setting, self-regulated learning does not entail teacher-structured tests, so feedback on monitoring can only be student-generated. Work by Dunlosky, Rawson, and colleagues (e.g., Dunlosky et al. 2011; Rawson and Dunlosky 2007) provides a framework for designing monitoring exercises without teacher feedback. Their research underlines the importance of having students generate and evaluate *pre-judgment knowledge retrieval* when studying key concepts in text books in order to improve monitoring accuracy and reduce overconfidence. In their studies, similar to everyday self-regulated learning, students read a set of texts containing key concepts and their definitions. One of these key concepts was, for example. "proactive interference" with the definition "Information already stored in memory interferes with the learning of new information". Prior to the students judging their knowledge of the key concept, they were required to generate the definition and evaluate its quality by comparing it to a standard, correct definition. Across several experiments (Dunlosky et al. 2011; Lipko et al. 2009) this retrieval-monitoring technique was shown to reduce overconfidence, particularly with respect to completely incorrect responses. Overconfidence on these 'commission errors' is particularly detrimental to learning, as students erroneously hold the belief that a completely incorrect response is (partially) correct. Commission errors are hypothesized to result from insufficient awareness of

the correct response, and actively comparing the generated definition to the correct definition helps to overcome this. However, even though overconfidence was reduced when comparing with the standard, students were still overconfident on about 30% of the commission errors. Perhaps comparing the self-generated definition to the standard was cognitively taxing and students were still not aware of the crucial elements that the definition should contain to be considered correct. In follow-up research (Dunlosky et al. 2011), key concept definitions were separated into idea units. As an example, the 'proactive interference' key concept contains three idea units: (1) 'Information already stored in memory', (2) 'interferes with the learning', and (3) 'of new information'. After prejudgment recall, students indicated which of these idea units were represented in the definition they provided. The idea-unit standard technique further reduced overconfidence on commission errors to about 10%. Most interestingly, this reduction was also observed when students themselves parsed the standard definition into idea units, which renders this technique an interesting option for improving students' self-regulated learning. That is, no teacher-feedback is needed and a clear effect on judgment accuracy is observed, which potentially affects final exam performance.

Inspired by the idea-unit standard technique, and with the ultimate goal to improve students' monitoring and regulation of learning during a college course to increase their exam performance, we designed an intervention that (1) would make students aware of the problem and presence of overconfidence during self-regulated learning of articles and text book chapters, and (2) would support them in monitoring and regulating their learning of articles and text book chapters more effectively. This intervention partly relied on the proven positive effects of self-testing, as the third step of the strategy consisted of actively retrieving studied materials from memory. Robust evidence shows that actively retrieving information from memory leads to higher memory than restudying information (Karpicke and Roediger 2007; Roediger and Karpicke 2006). Most (4 of the 7) steps of the strategy, however, focused on self-monitoring and regulating learning of the studied material. To separate the effects of these two parts of the intervention, we used a 2x2 factorial design exposing students to the awareness part of the intervention, the idea-unit technique, neither of these, or both. Contrary to previous work, the intervention was offered through the Internet to undergraduate college students of three different programs, to examine the potential of a content-free, cross-course transferrable intervention for improving monitoring and regulation of learning theoretical concepts.

Personality and overconfidence

A less emphasized, yet potentially important factor in research on monitoring accuracy in education is the role of personality traits. Given that overconfidence is known to affect learners of various ages and across a variety of tasks, it is possible that it covariates with specific personality characteristics. Metacognitive ability is suggested to be malleable and relatively independent of general intelligence (Pressley and Ghatala 1990; Veenman and Spaans 2005), but that does not exclude the possibility that certain personality traits correlate with monitoring accuracy and influence the degree of malleability. In laboratory-like learning environments, associations between personality and confidence judgments have been established. Although small, a positive relation between extraversion and overconfidence has been reported (Dahl et al. 2010; Pallier et al. 2002, & Schaefer et al. 2004). In a similar vein, there is evidence for an association between narcissism and overconfidence, possibly contingent on the mechanism by which people high in narcissism believe their intelligence is higher than actually measured (Campbell et al. 2004; Buratti et al. 2013). Buratti and colleagues

recently conducted an extensive study on the relation between monitoring accuracy, personality traits and cognitive style. They included a wide range of personality traits that potentially relate to metacognitive judgments and meta-metacognitive judgments (confidence judgments about confidence judgments) on a general knowledge test. Specifically, they hypothesized relations to Big-5 measures of extraversion, openness, conscientiousness, and agreeableness as well as self-doubt. Only openness and a combined factor of extraversion and narcissism related positively but to a small extent to overconfidence, whereas none of the cognitive style measures did. While the Buratti et al. study was inclusive in measures and types of judgments, its learning task could be considered to be low in ecological validity. Possibly, the general knowledge test was not fully experienced as an actual performance task, thereby altering the nature of the judgment task too. In this study, we therefore included the actual learning environment of the students and examined to what extent personality traits are related to monitoring accuracy. Moreover, previous research in this area neglected the distinction typically made between grandiose and vulnerable narcissism (Wink 1991), where grandiose narcissism is reflected in the overestimation of one's achievements, talents, and characteristics (Campbell et al. 2002; Carlson et al. 2011), whereas vulnerable narcissists also experience feelings of grandeur and have a high need for attention, but present themselves as shy and lacking self-confidence (Wink 1991). In an educational setting, both types of narcissism could contribute to overconfidence, although students high in vulnerable narcissism might be unconfident about their achievements in social situations. It is thus interesting to add this distinction and examine whether it adds to explaining overconfidence in education, and to scrutinize the relation between the overt insecurity vulnerable narcissists display and their covert overestimation.

Moreover, a cognitive style factor that has not yet received attention, but warrants investigation is optimism. Optimism refers to the extent to which people have positive expectations of the future (Carver et al. 2010). It is considered a stable trait that has been related to physical and psychological wellbeing (Carver et al. 2010). Optimism might however also have its 'side effects'. When optimism becomes unrealistic it is called optimistic bias. It entails the human tendency to expect more favorable outcomes for the future than could possibly be true (Shepperd et al. 2013). Unlike dispositional optimism, which is considered to be a trait, unrealistic optimism is a universal bias (Shepperd et al. 2013). It is thinkable that unrealistic optimism expresses itself as overconfidence in educational settings, and correlates with dispositional optimism. The effect of dispositional optimism on overconfidence in educational settings, however, has not been investigated so far. Dispositional optimism is usually measured by means of the Life Orientation Test – Revised (LOT-R, Carver et al. 2010).

The present study

In sum, the present study set out to answer three related research questions:

1) How does students' monitoring accuracy of exam grade predictions develop over the length of a college course?
2) Can overconfidence on exam grade predictions be reduced by an intervention designed to improve monitoring and regulation of learning and will this intervention affect exam grades too?
3) How do narcissism and optimism influence monitoring accuracy and development of monitoring accuracy during a college course?

To this end, first-year undergraduate students from three different faculties at a Dutch university were asked to predict their exam grades at the beginning as well as the end of a college course. They also completed questionnaires to measure their level of narcissism (both grandiose and vulnerable), and optimism. One week into the course, students were appointed to one of four conditions that received either (1) a monitoring exercise, (2) instruction on a monitoring and regulation strategy, (3) both, or (4) neither.

We hypothesized that overconfidence on exam grade predictions would decrease over the length of the college course for both high and poor performers, although poor performers were expected to still be overconfident at the end of the course. Moreover, we predicted that students who received either the monitoring exercise, the monitoring and regulation strategy, or both would on average have higher monitoring accuracy and a higher exam grade than students who did not receive an intervention. As the strategy provided concrete advice on how to improve self-regulated study, we expected students who received the strategy to outperform students who only received the monitoring exercise in both monitoring accuracy and their exam grade. With regard to the influence of personality factors, we hypothesized a positive relation between overconfidence and narcissism (both grandiose and vulnerable), and between overconfidence and optimism, on both measurement times.

Method

Sampling and participants

Students from 7 first-year courses in three faculties (Faculty of Psychology and Neuroscience, the Faculty of Health, Medicine and Life Sciences, or the School of Business and Economics) were invited to participate in the study during the introductory lecture of the course. A link to the questionnaires was provided in their electronic learning environment, where informed consent was also obtained.

Although 244 students signed the informed consent, a total of 226 students (153 female) actually participated in the study: Faculty of Psychology and Neuroscience: $N = 17$; Faculty of Health, Medicine and Life Sciences, $N = 118$, and School of Business and Economics, $N = 90$). One student was excluded because she was in a different curriculum and we could not obtain her final grade. One other student who did the self-test was excluded for the analysis of the Monitoring Exercise as she had already read the texts in another experiment. The mean age of the participants was 20.5 years ($SD = 1.87$). Because of the international character of the university, students could choose to participate in English or Dutch. One hundred twenty-two students participated in Dutch (87% Dutch nationality), and 103 students participated in English (14% Dutch nationality). Of the students participating in Dutch, 93% had Dutch as their first language. Of the students participating in English, 7% had English as their first language. Due to the online nature of the experiment, missing data and dropout occurred. A small number of participants filled in the questionnaires twice. In these cases, the first entry was used.

Design

Stratified by language, students were randomly assigned to one of four conditions: Monitoring Exercise (ME condition, $N = 76$), Monitoring and Regulation Strategy (MRS condition,

$N = 40$), both (ME-MRS condition, $N = 70$), or neither (control condition, $N = 39$). Approximately twice as many participants were assigned to the ME and ME-MRS conditions, in order to compensate for expected dropout. Number of participants who completed the conditions and provided an exam score prediction close to the test plus an actual exam score were $N = 44$ for ME condition, $N = 27$ for MRS condition, $N = 33$ for ME-MRS condition, and $N = 28$ for control condition.

Materials

Questionnaires and the intervention were delivered online, using the questionnaire tool Surveymonkey (www.surveymonkey.com). Number of participants who completed each questionnaire is mentioned below.

Monitoring exercise Similar to Rawson and Dunlosky's materials (2007), participants read two expository texts titled 'Gestures' (241 words) and 'Marriage' (264 words), each containing four key concepts and their definitions. See Appendix A for a sample text. As an example, the key concept 'vital marriage' was defined as "one in which partners have a deep emotional connection. Togetherness and sharing are very important and sex is viewed as a pleasant experience instead of an obligation". After reading both texts, participants were provided the key concepts and had to type a definition for each concept. In line with Dunlosky et al. (2005), participants then provided *pre-judgment retrieval*; While shown their own definition, they were asked to judge its quality (on a scale from 1 "completely incorrect" to 5 "completely correct"). This was done in absence of the correct definition. After giving the last judgment of the final key concept, the exercise was finished. At most 3 days after conducting the monitoring exercise, participants received an email containing feedback on their monitoring. This feedback stated how many of their definitions were incorrect, partially correct or fully correct, but also how many of their judgments were overestimations, underestimations or correct estimations of their performance.

Monitoring and regulation strategy Participants were provided with the monitoring and regulation strategy through e-mail. The complete two-page text of the strategy can be found in Appendix B. It outlines a 7-step procedure that explains students how to assess their knowledge of key terms in textbooks by self-testing and comparing their definitions with the correct definition through counting idea-units (referred to as 'keywords' within the definition) to come to accurate self-monitoring. It further emphasizes the importance of restudying key concepts that are not yet mastered.

Questionnaires

Vulnerable narcissism The Hypersensitive Narcissism Scale (HSNS, $n = 219$) was used to measure vulnerable or covert narcissism (Hendin and Cheek 1997). The questionnaire consists of 10 items. Example of an item is "I often interpret remarks of others in a personal way". The HSNS scale had a 5-point scale (1 = strongly disagree, 5 = strongly agree; all scale labels were named) in the English version but a 7-point scale in the Dutch version. We recoded both versions into a 0 to 1 scale, by changing the 1 into a 0, and the 5 and 7 into 1, and equally

distributing the remaining scale points between 0 and 1. A Cronbach's alpha of .73 was found for both the Dutch and the English version.

Grandiose narcissism The Narcissistic Personality Inventory (NPI, $n = 216$) is a commonly used questionnaire to measure characteristics of grandiose narcissistic personality in a healthy population (Raskin and Terry 1988). The NPI consists of 37 items that are scored on a 7-point scale (1 = strongly disagree, 7 = strongly agree; all scale labels were named). Example of an item is "I like to be the center of attention". We found a Cronbach's alpha of .89 for the Dutch version of the questionnaire and .91 for the English version.

Optimism The Life Orientation Test - Revised (LOT-R, $n = 218$) was used to measure generalized optimism (Scheier et al. 1994). The LOT-R consists of ten items, of which four items are filler items, measured on a 5-point scale (1 = I agree a lot, 5 = I disagree a lot; all scale labels named). Example of an item is "I am always optimistic about my future". Three of the other items are pessimistically formulated, whereas the other three are optimistically formulated. Cronbach's alpha was .73 for the Dutch version and .72 for the English version.

Procedure

The research was divided into three phases, a pre-intervention phase ($t = 0$), an intervention phase ($t = 1a$ and $t = 1b$) and a post-intervention phase ($t = 2$). See Table 1 for an overview of the experiment procedure.

The length of the course was 8 weeks. The factor 'time' was incorporated by comparing judgments between $t = 0$, $t = 1a$, and $t = 2$. Participants could access the questionnaires via a link in their electronic learning environment. The questionnaire asked for demographic information and a prediction of their course test grade on a scale from 0 to 10 ($t = 0$, $n = 225$), as is typical in the Netherlands. After this prediction, participants were presented with the NPI, HSNS, and LOT-R in random order. Then, participants were randomly assigned to one of four conditions. At most 3 days after finishing the questionnaire, participants in the ME-condition and the ME-MRS

Table 1 Overview of the experiment procedure per condition over time points during the course

	Time point	ME	MRS	ME-MRS	Control
$t = 0$	Start of the course	----All groups: Predicted exam score, Completed personality questionnaires----			
$t = 1a$	3 days later	Completed Monitoring Exercise		Completed Monitoring Exercise	
$t = 1b$	3 days after finishing Monitoring Exercise	Feedback on Monitoring Exercise		Feedback on Monitoring Exercise	
			Studied Monitoring and Regulation Strategy	Studied Monitoring and Regulation Strategy	
$t = 2$	3 days before exam	----------All groups: Predicted exam score----------			

ME Monitoring Exercise Group, *MRS* Monitoring and Regulation Strategy Group, *ME-MRS* Monitoring Exercise and Monitoring and Regulation Strategy Group

condition received an email with a link to the monitoring exercise ($t = 1$a). One week after this email, a reminder was sent to those who had not yet completed the exercise. At most 3 days after conducting the monitoring exercise, participants received the email containing feedback on their monitoring ($t = 1$b). Finally, participants were informed about the negative effect overconfidence would have on their self-regulated learning, and were urged to try to correctly estimate their own knowledge level. All participants but one overestimated their performance on at least one key concept (M nr of overestimations = 4.95, SD = 1.67, M nr of correct estimations = 1.44, SD = 1.21, M nr of underestimations = 1.56, SD = 1.48). Finally, participants were asked to complete a short questionnaire about the monitoring exercise and the feedback they received. The questionnaire assessed on a 1 (not at all) to 5 (very much) Likert scale how surprised they were about the result of the monitoring exercise, how insightful the exercise was, to what extent they thought the result indicated their true ability to assess their own learning, and to what extent they planned to change their way of studying because of the result of the monitoring exercise.

Again maximally 3 days after the monitoring exercise ($t = 1$b), participants in the MRS and ME-MRS conditions received a four-page guide that described a strategy for correctly monitoring and regulating learning by means of an example. Participants in the MRS condition were paired with a participant in the ME-MRS condition based on order of entering the experiment to equal average starting moment of the MRS intervention across conditions. When a participant in the ME-MRS condition had finished the monitoring exercise and received the monitoring and regulation strategy by email, the paired participant in the MRS condition also received the information on the regulation strategy by email. Participants were told to read the information carefully, and they were allowed to reread the strategy information as often as they wanted. Participants were also told to use the described strategy as often as possible. One week and 3 days before the exam, participants received an email and a text-message in which they were again asked to estimate their exam score ($t = 2$, $n = 132$). After the exam, all participants received a debriefing, including the strategy guide. Furthermore, 50 50-euro gift vouchers were raffled among the participants who completed all parts of the study.

Analyses

Absolute accuracy (Dunning et al. 2003) for exam score predictions at $t = 0$ and $t = 2$, as well as for the judgments during the monitoring exercise ($t = 1$a) was calculated. With regard to exam score predictions, absolute accuracy was determined by subtracting the actual exam score from the predicted exam score. Thus, positive values indicated overconfidence, whereas negative values referred to underconfidence. A score of zero indicated an accurate prediction. For both the predicted scores of key concept definitions and actual scores of key concept definitions, five points were awarded for correct responses, three points for partially correct responses and 1 point for incorrect responses. Again, for each key concept, the actual score was subtracted from the predicted score. The sum of this score for all 8 key concepts was a measure of absolute accuracy in the exercise. Cronbach's alpha of the key concept judgments was .63.

Data were analyzed using IBM SPSS Statistics 21 (IBM, Amsterdam, the Netherlands). An important concern in this study was the hierarchical structure of the data, because participants were nested within 7 different courses and took one of 7 different exams. Students in the same course could be expected to be more similar to each other compared to students in other courses (i.e. intra-course correlations). Multilevel analysis can be used to take this hierarchical structure between groups into account (Leppink 2015; Peugh 2010). Multilevel analysis distinguishes between fixed and random effects. Random effects refer to the courses being

considered a random sample of all courses in the population (i.e., we are not interested in the differences between courses per se, but we acknowledge that they explain variance in the sample). Therefore, a random intercept for course (with 7 levels) was included in the analyses, to take into account these intra-course correlations. Furthermore, a random intercept for 'language' (i.e., participated in the Dutch or English version) was included in each analysis. The fixed effects are the effects of interest, such as the intervention and personality variables. The −2 restricted log likelihood estimate was used as an information criterion to decide whether or not adding a random intercept yielded a significantly better fit to the data. If the random intercepts did not lead to a significantly better fit, these were not included in the analyses.

The effect of the intervention was estimated according to the intention to treat principle (Montori and Guyatt 2001) that classifies students based on their randomly assigned condition. This leads to less biased results, but due to possible non-compliance with the intervention, the results might be an underestimation of the actual effect.

Results

Dropout

Because of the online nature of the research and its embedding within actual university courses, there was dropout at several points. Two types of dropout were separately investigated: dropout from the intervention, and missing data at $t=2$. Dropout from the ME condition was defined as a participant who did not complete the monitoring exercise (dropout 24 from 76 participants). Dropout from the MRS condition was defined as a participant who did not fill in the questionnaire about the strategy (dropout 14 from 40 participants). Dropout from the ME-MRS condition was defined as a participant who either did not do the monitoring exercise, or did not fill in the questionnaire about the strategy (or did neither) (dropout 39 from 70 participants). Dropout from the control condition was not possible, because there was no intervention. Missing data at $t=2$ occurred when participants did not provide a second exam score prediction (dropout 96 from 225 participants).

To investigate the nature of the dropout across conditions, a logistic regression analysis was conducted with dropout (yes/no) as a dependent variable and exam score, absolute accuracy at $t=0$ and $t=2$, sex, language, course and condition as independent variables. None of these variables significantly predicted dropout from the intervention (all $ps>.09$). A second dropout analysis was conducted to investigate dropout at $t=2$ (i.e., participants who did not provide a second prediction of their exam score). Optimism and actual exam score were significant predictors for not providing a second grade prediction. Scoring higher on optimism increased chance of not providing an exam score prediction ($b=.09$, SE = .04, $p=.02$), and having a lower exam score increased chances of not providing an exam score prediction ($b=-.35$, SE = .14, $p=.01$) .

Time and absolute accuracy

To investigate how students' monitoring accuracy of exam grade prediction developed over time for students of different performance levels, participants were divided in quartiles based on their actual exam score, with Q1 being the lowest quartile (See Table 2). Language and

Table 2 Mean exam score predictions at the start ($t = 0$) of the course and close to the exam ($t = 2$) for each quartile of performers

	Quartile	1	2	3	4
At $t = 0$	Exam Score Predictions	6.42	6.89	6.70	7.32
	Actual Exam Score	4.64	6.00	6.83	8.12
	Difference	1.79	0.89	−0.13	−0.80
At $t = 2$	Exam Score Predictions	5.79	6.36	6.63	7.09
	Actual Exam Score	4.64	6.00	6.83	8.12
	Difference	1.15	0.36	−0.20	−1.03

Q1 = exam scores of the 1st quartile = lowest grades, Q2 = exam scores of the 2nd quartile, Q3 = exam scores of the 3rd quartile, Q4 = exam scores of the 4th quartile = highest grades

course were redundant as random factors, so a 2 × 4 repeated measures ANOVA was conducted with absolute accuracy at $t = 0$ and $t = 2$ ('time') as the within-subjects factor and performance quartile group on the exam ('quartile') as the between-subjects factor (see Table 4). There was a significant main effect of time, $F(1, 125) = 38.15$, $p < .001$, $\eta_p^2 = .234$, indicating that the difference between predictions and exam scores decreased from $t = 0$ to $t = 2$. A significant main effect of quartile was also found, $F(3, 125) = 49.06$, $p < .001$, $\eta_p^2 = .54$, indicating that all quartiles differed in their absolute accuracy, with the two lowest quartiles being overconfident and the highest quartile being underconfident. The interaction between time and quartile was not significant, $F(3, 125) = 1.79$, $p = .15$, $\eta_p^2 = .04$, so the difference between quartile groups in absolute accuracy did not change over time. At the start of the course, 91.1% of the students (41 out of 45 students) who ultimately failed the exam believed they would pass it (i.e., predicted to obtain a score of at least 5.5, the minimum pass grade, on a 10-point scale). Most alarming, 3 days before the exam still 71.4% of those who failed (15 out of 21 students. Note: N differs due to drop-out at $t = 2$) believed they would pass the exam.

Effect of the monitoring exercise and monitoring and regulation strategy

Effect on absolute accuracy A 2 × 2 ANCOVA was conducted with absolute accuracy at $t = 2$ as a dependent variable, and absolute accuracy at $t = 0$ as a covariate. Between-subjects factors were Monitoring Exercise (yes/no) and Monitoring and Regulation Strategy (yes/no). The covariate, absolute accuracy at $t = 0$, was significantly related to absolute accuracy at $t = 2$, $F(1,124) = 213.241$, $p < .001$. Absolute accuracy at the start of the course clearly predicted absolute accuracy close to the exam date. The main effect of ME was not significant, $F(1,124) = 2.29$, $p = .133$, and neither was the effect of MRS, $F(1,124) = .034$, $p = .855$. The interaction between ME and MRS was approaching significance, $F(1,124) = 3.19$, $p = .077$ (See Table 3 and Figure 1).

Students who received the monitoring and regulation strategy were marginally significantly more accurate in their exam score predictions than those who did not receive the strategy. Those who only did the Monitoring Exercise were underconfident, whereas those in the control group were overconfident. These students who did not get any intervention overestimated their performance with almost half a point at $t = 2$ (M = .45, SD = 1.1). This mean differed significantly from zero in a one-sample t-test, $t(27) = 2.17$, $p = .039$. Absolute accuracy in the three experimental groups was not significantly different from zero

Table 3 Absolute accuracy at the start ($t = 0$) of the course, close to the exam ($t = 2$), and exam score for each condition. Standard deviations between brackets

		Control	ME	MRS	ME-MRS
$T = 0$	Exam score predictions	7.20 (0.75)	6.76 (0.97)	6.85 (0.82)	7.03 (0.88)
	Absolute accuracy	0.71 (1.44)	0.22 (1.32)	0.07 (1.13)	0.23 (1.01)
$T = 2$	Exam score predictions	6.93 (0.73)	6.21 (1.14)	6.51 (0.87)	6.69 (0.92)
	Absolute accuracy	0.45 (1.11)	−0.33 (1.27)	−0.27 (1.09)	−0.12 (1.06)
	Exam score	6.48 (1.32)	6.54 (1.38)	6.78 (1.25)	6.81 (1.16)

Negative absolute accuracy indicates underconfidence while positive absolute accuracy indicates overconfidence

ME Monitoring Exercise Group, *MRS* Monitoring and Regulation Strategy Group, *ME-MRS* Monitoring Exercise and Monitoring and Regulations Strategy Group

($M_{ME} = -.33$, $SD_{ME} = 1.27$, $t(42) = -1.68$, $p = .10$; $M_{MRS} = -.27$, $SD_{MRS} = 1.09$, $t(26) = -.267$, $p = .22$; $M_{ME\text{-}MRS} = -.12$, $SD_{ME\text{-}MRS} = 1.06$, $t(30) = -.63$, $p = .53$). Thus, no significant amount of overconfidence was observed in any of the experimental groups close to the exam.

Effect on exam score Mean (SD) exam scores per course are provided in Table 4.

Including the random factors 'course' and 'language' in the analysis yielded a significantly better fit $\chi 2(2) = 17.353$, $p < .001$, so a multi-level analysis was conducted with exam score as a dependent variable, and students' grade point average (GPA) as a predictor. Between-subjects factors were Monitoring Exercise (yes/no) and Monitoring and Regulation Strategy (yes/no). Results revealed that GPA was a significant predictor of exam score, $b = .88$, $SE = .05$, $t(193.93) = 16.66$, $p < .001$. Since the interaction of the two between-subjects factors was not significant, $b = -.23$, $SE = .25$, $t(190.78) = -.93$, $p = .35$, it was removed from the model for reasons of parsimony. Moreover, a main effect of MRS was found, $b = -.26$, $SE = .12$, $t(193.32) = -2.25$, $p = .025$, but no main effect of ME, $b = .14$, $SE = .12$, $t(192.00) = -1.17$, $p = .24$. Apparently, the Monitoring and Regulation Strategy affected students' study behavior resulting in an increased exam score, but the Monitoring Exercise did not (neither in isolation nor in combination with the MRS). Participants in the control condition scored on average 6.35 on a

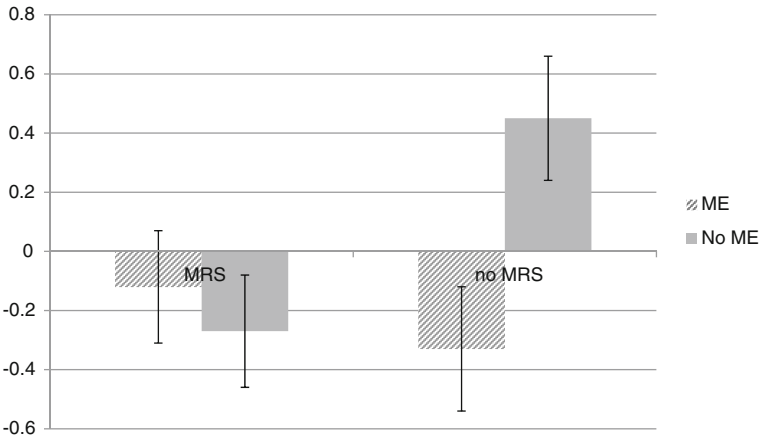

Fig. 1 Mean absolute accuracy of exam score prediction at $t = 2$ for the Monitoring Exercise (ME) and Monitoring and Regulation Strategy (MRS) intervention. *Error bars* denote standard error

Table 4 Exam scores per course (course topic provided, means, SDs between brackets)

Faculty		Exam score
Faculty of Health, Medicine, and Life Sciences	Genetics ($n = 30$)	6.63 (1.00)
	Aging ($n = 6$)	6.17 (1.47)
	Internal medicine ($n = 45$)	6.38 (1.34)
	Care ($n = 28$)	7.11 (1.23)
	Internal medicine (international) ($n = 5$)	5.00 (1.58)
Faculty of Psychology	History of psychology ($n = 16$)	6.09 (1.43)
School of Business and Economics	Finance ($n = 87$)	6.31 (1.30)

Course exams are graded on a scale from 1 (minimum score) to 10 (maximum score)

10-point scale ($SD = 1.32$). In the ME condition, the average exam score was also 6.35 ($SD = 1.31$). In the MRS condition, the mean exam score was 6.78 ($SD = 1.28$) and in the ME-MRS condition, the mean exam score was 6.42 ($SD = 1.32$). Table 5 presents students' evaluation of the monitoring exercise, no significant differences were found between conditions, all ts < 1.

Overconfidence & personality

The relationship between absolute accuracy, grandiose (NPI) and vulnerable narcissism (HSNS), and optimism (LOT-R) was investigated in two multiple regression analyses at $t = 0$ and $t = 2$ (See Table 6). In both analyses, language and sex were included as fixed factors. Furthermore, at $t = 0$, a random intercept for 'course' was included, making this model a multilevel analysis. This model had a significantly better fit than the model without the random intercept, $\chi 2(1) = 7.458$, $p = .0063$. For the analysis at $t = 2$, no random intercept was included, as this did not lead to a better fit of the model. The results at $t = 0$ show that students who have more characteristics of grandiose narcissism (a higher NPI score) showed more overestimation of their exam score, whereas students with more characteristics of vulnerable narcissism (a higher HSNS score) showed less overestimation of their exam score. Furthermore, the LOT-R was not significantly related to absolute accuracy. The language effect indicates that students who participated in English were significantly more overconfident than students participating in Dutch.

The same analysis was run for absolute accuracy at $t = 2$ (See Table 6). At $t = 2$, again, the NPI score was significantly positively related to absolute accuracy. However, the HSNS score was not significantly related to absolute accuracy of the exam score predictions close to the exam date. LOT-R scores were also not significantly related to absolute accuracy.

Table 5 Evaluation of monitoring exercise by students per group on a scale from 1 (not at all) to 5 (very much so) (Means, SDs between brackets)

	MRS group	ME-MRS group
Understandability	4.31 (0.55)	4.41 (0.76)
Insightfulness	3.54 (0.81)	3.59 (0.87)
Usefulness	3.15 (0.92)	3.06 (0.76)
Intention to use	2.62 (0.57)	2.44 (0.88)

MRS Monitoring and Regulation Strategy, *ME-MRS* Monitoring Exercise and Monitoring and Regulation Strategy

Table 6 Results of analyses with absolute accuracy at $t = 0$ and $t = 2$ as dependent variables and personality measures as independent variables

Parameter	Absolute accuracy at $t = 0$					
	B	CI	SE	df	t	p
Intercept	0.402	[−0.881, 1.685]	0.650	165.985	0.619	0.537
Language = English	−0.574	[−1.100, −0.049]	0.263	59.354	−2.187	0.033
Sex = female	−0.307	[−0.661, 0.047]	0.179	198.850	−1.712	0.088
HSNS	−0.153	[−0.265, −0.040]	0.057	198.973	−2.682	0.008
LOT-R	−0.010	[−0.047, 0.026]	0.019	196.752	−0.563	0.574
NPI	0.011	[0.004, 0.017]	0.003	196.923	3.267	0.001
	Absolute accuracy at $t = 2$					
	B	CI	SE	df	t	p
Intercept	−1.239	[−2.771, 0,293]	0.774	122	−1.601	0.112
Language = English	−0.555	[−1.057, −0.053]	0.254	122	−2.189	0.030
Sex = female	−0.094	[−0.565, 0.378]	0.238	122	−0.394	0.695
HSNS	−0.028	[−0.175, 0.119]	0.074	122	−0.377	0.707
LOT-R	0.006	[−0.041, 0.053]	0.024	122	0.248	0.805
NPI	0.011	[0.0023, 0.019]	0.004	122	2.686	0.008

HSNS Hypersensitive Narcissism Scale, *LOT-R* Life Orientation Test – Revised, *NPI* Narcissistic Personality Inventory

We also analyzed the relation between the personality factors and absolute accuracy for their key term definitions during the Monitoring Exercise ($t = 1$a) in a multiple regression (See Table 7). Again, language and sex were included as fixed factors. Male students were marginally significantly more overconfident when estimating the quality of their key term definitions. No effect of language was found here. The three personality measures (LOT-R, NPI, HSNS) were also not significantly related with absolute accuracy during the Monitoring Exercise. Note that this analysis was based on a rather small number of participants.

Moreover, there was no significant correlation between absolute accuracy at $t = 0$ and absolute accuracy during the Monitoring Exercise, $r = .15$, $p = .18$, showing that these measurements of monitoring accuracy are based on (partially) different cues (Table 8).

Table 7 Results of multiple regression analysis with absolute accuracy during monitoring exercise ($t = 1$) as dependent variables and personality measures as independent variables

Parameter	Absolute accuracy during Monitoring Exercise ($t = 1$)			
	B	SE	t	p
Intercept	−3.68	4.73	-.78	.44
Language = English	1.87	1.46	−1.29	.20
Sex = female	−2.33*	1.27	−1.84	.07
HSNS	.75	.46	1.64	.11
LOT-R	.02	.13	.17	.87
NPI	.03	.02	1.40	.17

HSNS Hypersensitive Narcissism Scale, *LOT-R* Life Orientation Test – Revised, *NPI* Narcissistic Personality Inventory

*$p = .069$

Table 8 Simple correlations between the personality questionnaires, absolute accuracy at $t = 0$ and $t = 2$, and exam score predictions at $t = 0$ and $t = 2$

	HSNS		LOT-R		NPI	
	r	p	r	p	r	p
Absolute Accuracy $t = 0$	−0.154*	0.025	-.181**	0.009	.263**	<0.001
Absolute Accuracy $t = 2$	−0.003	0.974	−0.127	0.152	.306**	<0.001
Exam score prediction $t = 0$	−0.130	0.055	−0.132	0.052	.174*	0.011
Exam score prediction $t = 2$	0.038	0.664	0.024	0.787	.215*	0.013

HSNS Hypersensitive Narcissism Scale, *LOT-R* Life Orientation Test – Revised, *NPI* Narcissistic Personality Inventory

* $p < .05$

** $p < .001$

Discussion

A pivotal step in self-regulated learning is accurate self-monitoring of learning. Little work has been done to examine how in real higher-education environments self-monitoring can be trained. The present study examined how feedback on monitoring accuracy through a monitoring exercise and a monitoring and regulation strategy could improve first year college students' monitoring of learning and contribute to academic achievement. Moreover, we analyzed how monitoring accuracy changed from the start to the end of the course, and how personality factors as narcissism and optimism influenced monitoring accuracy.

Starting with the development of monitoring accuracy over the duration of a college course, we found that there was a linear decrease in exam score predictions across all performance quartiles from the start to the end of the course. This led poor performers to become less overconfident and highest performers to become somewhat more underconfident. Three days before the exam, more than 70% of those who failed the exam did not expect to fail. Replicating previous research (e.g., Dunning et al. 2003), the performance quartiles differed in absolute accuracy with the greatest discrepancy between prediction and exam score in the lowest quartile. These findings show that knowledge plays a role in exam score predictions, given the decrease of overconfidence as knowledge increases over a college course. What these findings add is that knowledge is not enough to explain the typical over/underconfidence pattern first observed by Dunning and colleagues (2003). If knowledge level plays a crucial role, an interaction between absolute accuracy and performance quartile would have been observed, with high performers having highly accurate exam score predictions and poor performers still being overconfident. The observation that high performers were more underconfident at the end compared to the start of the course and the resulting lack of an interaction demonstrate that metacognitive cues used also plays a role. Possibly, high performers derive different experience-based cues from their knowledge gain by making them overly aware of their potential knowledge gaps, whereas poor performers are still not fully aware of their knowledge gaps by the end of the course. From an information-based cue perspective, it is possible that poor performers make a wishful thinking prediction at the start of the course and slightly but insufficiently adjust this prediction towards the end of the course, whereas the highest performers are too cautious in their predictions at both times. The current design did not allow for measurement of actual cue use, and thus we are unable to determine to

what extent information-based cues, experience-based cues or both were affected through our intervention (Koriat et al. 2008). Further research into cue use when making exam score predictions, potentially including analyses of students' explanations for exam score predictions are needed to shed light on this issue and tally personalized interventions for the different performance groups. Note, though, that asking students to explain their exam score predictions will provide only partial insight as not all cue use is explicit and conscious (Nisbett and Wilson 1977). Analyzing exam score predictions over time, however, shows that the measurement artifact explanation (Krueger and Mueller 2002) is unlikely: Under that explanation no difference of monitoring accuracy over time would have occurred. It also shows that metacognitive ability is indeed malleable (Pressley and Ghatala 1990; Veenman and Spaans 2005), and thus sensitive to intervention. Note that little is known about what explains underconfidence in education and to what extent it harms self-regulated learning. Test anxiety, for one, does not appear to be related (Miesner and Maki 2007). Future research should look into cue use related to *under*confidence on exam score predictions. Given that the judgments were collected at a global test score level instead of the test item level, we are unable to determine whether a hard-easy effect occurred as this is typically observed when averaging item level judgments (Merkle 2009). However, given that we observed overconfidence even at the global test score level, we can conclude that this pattern is distinct from the typical item-level hard-easy effect and warrants further investigation.

The second aim of this study was to examine how higher education students' monitoring and regulation of learning could be improved through an online, self-study intervention. Students either received feedback on their monitoring accuracy through a Monitoring Exercise (based on studying two short texts), studied a Monitoring and Regulation Strategy, received both or neither of these. Our findings revealed an interaction effect on absolute accuracy that was approaching significance. This can be cautiously interpreted as indicating that exam score predictions in the groups that received either or both the Monitoring Exercise and the Monitoring and Regulation Strategy became closer to actual exam scores compared to the control group. Moreover, those who received both or one part of the intervention increased absolute accuracy to near zero. Apparently, receiving feedback on monitoring accuracy or studying a strategy to monitor, regulate and self-test knowledge positively affected students' estimations of their exam scores. The Monitoring and Regulation Strategy also affected their actual learning as observed through increased exam scores. Students receiving the strategy possibly made their self-study more effective. To what extent that was due to one or a combination of the three components of the Strategy (monitoring, regulation, or self-testing) is a question for future research in which these components are experimentally varied. Note that self-testing (Karpicke and Roediger 2007; Roediger and Karpicke 2006) was part of only one step (Step 3) of the strategy whereas monitoring and regulation were central to Step 4, 5, 6, and 7. Therefore, and given the effect the MRS had on absolute accuracy, we consider it unlikely that the effect on exam score was solely due to self-testing.

Combining the findings on absolute accuracy and exam score leads us to conclude that the Monitoring and Regulation Strategy was more effective than the Monitoring Exercise in improving monitoring and regulation of learning in this set of higher education students. Even though there was no guided practice with the strategy and no logging of strategy use was performed, students were able to benefit from processing the strategy, which resulted in improved academic performance. Future research should concentrate on examining how students can effectively implement the strategy in their self-regulated learning activities and study what aspects of the strategy are most effective in improving monitoring accuracy and

academic performance. The observation that the Monitoring Exercise alone did not solicit an effect on the exam score is not unexpected: the exercise focused solely on making students aware of the inaccuracy of their knowledge predictions and did not provide hands-on advice on how to improve self-regulated learning. More surprising is that the Monitoring and Regulation Strategy had a similar effect with or without the Monitoring Exercise. Apparently, an initial confrontation with inaccurate monitoring is not essential to lead students to adapting their self-regulated learning through a strategy.

Our third research question on the influence of personality traits on overconfidence revealed a modest effect of narcissism, although only for the more 'global' exam score predictions, and no effect of optimism. This study is the first to differentiate between grandiose and vulnerable narcissism in relation to overconfidence in education: Students higher in grandiose narcissism showed more overconfidence, whereas those higher in vulnerable narcissism showed *less* overconfidence. Contrary to our expectations, high vulnerable narcissism seems to protect against overconfidence in exam scores. These findings indicate that differentiating between grandiose and vulnerable narcissism is relevant in future research to come up with individualized interventions. Also contrary to our expectations, optimism did not contribute to students' monitoring accuracy. Apparently, a more or less optimistic cognitive style does not translate into predictions of exam scores. Overconfidence on exam score predictions therefore does not seem to relate to high dispositional optimism. Given that higher optimism was related to more drop out, it is possible that the lack of a relation between optimism and exam score predictions is due to a restricted range in optimism. However, we render this explanation unlikely: At the start of the course, when no drop out was yet observed, optimism did not influence absolute accuracy either.

The observation that personality traits did not influence the in-the-moment judgments of key term definitions in the Monitoring Exercise seems to indicate that these judgments are more affected by contextual and cognitive factors than by students' personality. The marginally significant effect of gender in this analysis should be followed up on in future research. Note, though, that both males and females were overconfident on the Monitoring Exercise. Finally, we observed that students who participated in English (14% Dutch nationality, 7% English as first language) showed poorer monitoring accuracy than those who participated in Dutch (87% Dutch nationality, 93% Dutch as first language). This difference could be due to students participating in English having lower knowledge of the Dutch educational and testing system when making exam score predictions, or to their processing the information mostly in their second language. Future research should attempt to disentangle the contribution of each or both of these factors.

Limitations and implications

While the real-life setting of this study allowed us to directly test the feasibility and impact of providing students with a monitoring and regulation intervention on their academic performance, the resulting lack of control over all measured variables leads to specific limitations. First, the limited number of participants, although considerable, prevented us from examining how personality traits and the intervention might have interacted. Even though there is little reason to assume that the Monitoring and Regulation Strategy might work less for individuals scoring high on narcissism, this is an empirical question that needs answering. Moreover, the current set-up of the

study did not allow for logging of self-regulated learning activities of students (such as actual use of the strategy), which would be needed to examine the impact the Monitoring and Regulation Strategy has on students' self-study activities. In-depth analysis could contribute to refining the strategy and design of practice exercises to maximize its effectiveness. Also, because of ethical considerations, we were unable to take into account actual exam difficulty. The difficulty of the exams potentially influenced absolute accuracy. Future research should incorporate exam difficulty, for instance parallel to research by Ackerman and colleagues (Ackerman et al. 2016) who manipulated task difficulty and showed how it influenced accuracy of confidence judgments in usability testing. Finally, this study was limited to first-year undergraduates of three faculties in one university. Up scaling it to different years, schools, and universities will add to the generalizability of these findings.

The real-life learning setting of our study does allow for outlining a number of practical implications. First, the widespread overconfidence on exam score predictions across the three faculties should be a concern for both students and universities. We know of no universities that formally train monitoring accuracy and regulation of learning in students, while this study provides evidence that this is possible even through a short online intervention. However, our findings underline that such an intervention should focus on providing information on and practicing with a Monitoring and Regulation Strategy instead of solely making students aware of their poor monitoring accuracy. Based on the current findings, we conclude that the latter does not even appear necessary. We here see a clear role of online learning that provides students continuous feedback on monitoring and regulation of learning while students engage in self-regulated learning activities. This study is just one in a short list (e.g., Kleitman and Costa 2014; Miller and Geraci 2011a; Nietfeld et al. 2006), which indicates the lack of attention for this theme in higher education research. However, it also proves that possible solutions are becoming tangible and warrant further exploration.

Compliance with ethical standards

Funding This research was funded by the Netherlands Organisation for Scientific Research (grant nr. 451-10-035) and the Maastricht University Leading in Learning Fund.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A

Marriage

Depending on individual preference and socioeconomic background, some people choose to highlight either the practical (utilitarian) or emotional (intrinsic) benefits of marriage. Scientists have distinguished four types of marriage relationship; two of which lean towards the utilitarian and two of which lean towards the intrinsic benefits. A DEVITALISED MARRIAGE is a marriage in which the initial passion, intimacy and companionship gives way to a utilitarian relationship. The partners spend little time together, experience little joy in sex and share few interests and activities. The majority of the time that they do spend together is filled with obligations, such as raising the children. A PASSIVE-CONGENIAL MARRIAGE is a

utilitarian relationship in which the partners have emphasised characteristics other than emotional closeness from the very start, unlike devitalised partners. Passive-congenial partners never expected marriage to bring emotional intensity and stress the practicality of their decision to get married. These couples stress the importance of their civil and professional responsibilities, their possessions, their economic security and childrearing. A VITAL MARRIAGE, on the other hand, is one in which the partners have a deep emotional connection. Togetherness and sharing are very important and sex is viewed as a pleasant experience instead of an obligation. TOTAL MARRIAGE resembles a vital marriage in that partners feel a deep emotional bond, but they share more aspects of their lives. The partners have similar careers and share the same friends and hobbies. The researchers stress that this classification only represents different perceptions on married life and says nothing about marital satisfaction; there are happy and unhappy couples in each of these categories.

Appendix B

A strategy to assess your knowledge

Many students find it difficult to determine whether they know the study material well enough. Figuring this out is important for two reasons: if you know the material well you can stop studying, whereas if you don't know it well enough you have to continue studying. Being prepared means you won't have to face any unpleasant surprises during the exam.

What is the most effective way to learn dozens of new terms and their definitions? How do you know what to focus on and what needs to be restudied? This document will teach you how to test your knowledge of study material. You may already be applying some of these techniques, but it's important to work through each step anyway. These steps have been scientifically proven to be effective.

Step 1 Study

Study the entire chapter as you normally would and write down all key terms on a separate piece of paper.

Step 2 Take a break!

Take a 10 min break before testing your knowledge. Make a cup of tea, browse through Facebook or have a chat with your housemate. This will put your mind on something else and make the following steps more effective.

Step 3 Provide definitions of key terms

Take your list of key terms and provide a definition for each one (don't cheat!).

Step 4 Assess your knowledge

Ask yourself how accurate your definition of each term is. Give yourself a score of 0 (incorrect), 0.5 (partially correct) or 1 (fully correct) per definition. Only give yourself a half credit or a whole

point if you think you would get the same score in the exam. Add up your credits and divide the total by the number of definitions. What percentage of the material do you think you know?

Step 5 Evaluate your assessment

Look at your definitions. Most textbooks have a glossary at the back, but if yours doesn't then open your book at the relevant chapter. Compare your definition to the one in the book and highlight all of the keywords within the definition. Do the same with your own answer and then compare how many keywords from the book's definition are in your own definition.

Give yourself 0, 0.5 or 1 credit for each definition. Only gives yourself 1 credit if you included all of the information from the book in your answer. Only give yourself half a credit if you included at least half of the information from the book in your answer. When you're done, add up your points and divide the total by the number of definitions. What is your estimated exam grade now?

Step 6 Compare

Compare your answers from steps 4 and 5. Did you overestimate your knowledge, underestimate it or have it spot on?

Step 7 Study more

Which terms did you give 0 points and half a point to? Study them again!

Take-home message

Step 5 will take you quite some time, is that really necessary? Yes, it is! Many problems arise because students award themselves too many points in step 4. They overestimate their own knowledge, do not study enough, and get into trouble during the test. In order to assess your own knowledge, and thus estimate your own grade, it is crucial to compare your own answers with the book. So rely on your comparison with the book (step 5), and not on your own assessment (step 4). Good luck!

References

Ackerman, R., Parush, A., Nassar, F., & Shtub, A. (2016). Metacognition and system usability: incorporating metacognitive research paradigm into usability testing. *Computers in Human Behavior, 54*, 101–113.

Bol, L., Hacker, D., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal of Experimental Education, 73*, 269–290.

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: the influence of personality traits and cognitive styles. *Metacognition and Learning, 8*(1), 79–102.

Campbell, W. K., Rudich, E., & Sedikides, C. (2002). Narcissism, self- esteem, and the positivity of self-views: two portraits of self-love. *Personality and Social Psychology Bulletin, 28*, 358–368.
Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making, 17*(4), 297–311.
Carlson, E. N., Vazire, S., & Oltmanns, T. F. (2011). You probably think this paper's about you: narcissists' perceptions of their personality and reputation. *Journal of Personality and Social Psychology, 101*, 185–201.
Carver, C. S., Scheier, M. F., & Segerstrom, S. C. (2010). Optimism. *Clinical Psychology Review, 30*, 879–889.
Dahl, M., Allwood, C. M., Rennemark, M., & Hagberg, B. (2010). The relation between personality and the realism in confidence judgements in older adults. *European Journal of Ageing, 7*(4), 283–291.
Dunlosky, J., Rawson, K. A., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52,* 551–565.
Dunlosky, J., Hartwig, M., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64*, 467e484. doi:10.1080/17470218.2010.502239.
Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83–87.
Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. doi:10.1111/j.1529-1006.2004.00018.x.
Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.
Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. doi:10.1037/0022-0663.92.1.160.
Händel, M., & Fritzsche, E. S. (2015). Students' confidence in their performance judgements: a comparison of different response scales. *Educational Psychology, 35*, 377–395.
Hartwig, M., & Dunlosky, J. (2014). The contribution of judgment scale to the unskilled-and-unaware phenomenon: how evaluating others can exaggerate over- (and under-) confidence. *Memory & Cognition, 42*, 164–173.
Hendin, H. M., & Cheek, J. M. (1997). Assessing hypersensitive narcissism: a reexamination of Murray's Narcissism Scale. *Journal of Research in Personality, 31*(4), 588–599. doi:10.1006/jrpe.1997.2204.
Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.
Kelemen, W. L., Winningham, R. G., & Weaver, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*(4–5), 689–717.
Kleitman, S., & Costa, D. S. J. (2014). The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance. *Learning and Individual Differences, 29,* 150–161.
Koriat, A. (1997). Monitoring one's knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.
Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 117–135). New York: Psychology Press.
Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*(2), 180–188.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.
Leppink, J. (2015). Data analysis in medical education research: a multilevel perspective. *Perspectives on Medical Education, 4*, 14–24.
Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities: The state of the art up to 1980. In D. Kahneman, P. Slovjc, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.
Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle-school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*, 307–318. doi:10.1037/a0017599.
Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychonomic Bulletin & Review, 16*, 204–213.
Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology, 19*, 650–670.
Miller, T. M., & Geraci, L. (2011a). Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314.

Miller, T. M., & Geraci, L. (2011b). Unskilled but aware: reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 502–506. doi:10.1037/a0021802.

Montori, V. M., & Guyatt, G. H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal, 165*(10), 1339–1341.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159–179.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Pallier, G., Wilkinson, R., Danthiir, V., Kleintman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*(3), 257–299.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85–112.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29*, 62–67.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self- regulated learning. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43–97). Lincoln: Buros Institute of Mental Measurements.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: monitoring learning from text. *Educational Psychologist, 25*, 19–33.

Price, P. C. (1998). Effects of relative-frequency elicitation question on likelihood judgment accuracy: the case of external correspondence. *Organizational Behavior and Human Decision Processes, 76*, 277–297.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54*(5), 890–902. doi:10.1037/0022-3514.54.5.890.

Rawson, K., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*, 559–579. doi:10.1080/09541440701326022.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*, 249–256.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the Big Five. *Journal of Research in Personality, 38*(5), 473–480. doi:10.1016/j.jrp.2003.09.010.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*(6), 1063–1078. doi:10.1037/0022-3514.67.6.1063.

Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8*(4), 395–411. doi:10.1177/1745691613485247.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66e73. doi:10.1037/0022-0663.95.1.66.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: age and task differences. *Learning and Individual Differences, 15*, 159–176.

Wink, P. (1991). 2 faces of narcissism. *Journal of Personality and Social Psychology, 61*(4), 590–597. doi:10.1037//0022-3514.61.4.590.

Zimmerman, B. J. (2000). Self-efficacy: an essential motive to learn. *Contemporary Educational Psychology, 25*, 82–91.